DS1004 FINAL REPORT

NEW YORK UNIVERSITY, CENTER FOR DATA SCIENCE

AN ANALYSIS TO THE EFFECT OF SOCIAL BEHAVIOR AND WEATHER CONDITIONS ON CRIME RATE IN NYC

May 10, 2017

Yichao Chen, yc1228 | Zoe Ma, ym910 | Yanjia Zhang, yz855

Abstract

We closely examine the NYPD crime dataset with external social behavior and weather dataset that we think relevant to crime rate. We use Pyspark on the HDFS to minimize our computational cost, SQL to format, select and join large relations, and Plotly package in Python to visualize our results. We formulate 6 hypotheses that we think may correlate with crime rate, including poverty rate, unemployment rate, graduation rate, heavy drinking youth rate, number of heavy drinker seeking medical aid, and weather, and verify using time series, correlation and regression analysis.

Contents

1 Introduction **3**

2 PART I: Data Summary and quality issues **3**

2.1 Data Description 3

2.2 Data quality 3

2.3 Detailed quality issues and cleaning 5

2.4 Data summary 7

3 PART II: Data Exploration **8**

3.1 Experimental techniques and methods 8

3.2 Hypothesis I: Poverty rate is positively correlated with crime count 10

3.3 Hypothesis 2: Unemployment rate has a leading effect on crime count 12

3.4 Hypothesis 3: Emphasis on education will decrease the crime rate 13

3.5 Hypothesis 4: High youth heavy drinking rate is a strong indicator for high crime rate 14

3.6 Hypothesis 5: The number of heavy drinkers seeking for medical support is negatively correlated to crime count 14

3.7 Hypothesis 6: Temperature can influence the number of happening of some specific crime types 15

4 Contribution **19**

5 Conclusion **19**

References **20**

1 INTRODUCTION

Crime rate in NYC has continuously dropped during recent years. Understanding the crime pattern and why the rate is dropping has important social impact for both the government and citizens. In this project, we research the crime patterns over all five boroughs in New York City, and analyze the social behaviors variables, which may result in the change on crime rate.

2 PART I: DATA SUMMARY AND QUALITY ISSUES

2.1 Data Description

The NYPD Complaint Data Historic dataset includes all valid felony, misdemeanor, and violation crimes reported to NYPD from 2006 to 2015. The original dataset contains 510,1231 rows of entries and 24 columns of features, including: complaint id `CMPLNT_NUM`, the exact date and time of occurrence of reported event `CMPLNT_FR_DT, CMPLNT_FR_TM`, the ending date and time of the reported event if applicable `CMPLNT_TO_DT, CMPLNT_TO_TM`, the report date of the crime `RPT_DT`, a three digit offense classification code and corresponding description `KY_CD, OFNS_DESC`, a more detailed classification code and description `PD_CD, PD_DESC`, an indicator of whether the crime was attempted but failed `CRM_ATPT_CPTD_CD`, an indicator of the level of offense `LAW_CAT_CD`, the Jurisdiction responsible for incident `JURIS_DESC`, the name of the borough `BORO_NM`, the precinct `ADDR_PCT_CD`, specific location and description of premises `LOC_OF_OCCUR_DESC, PREM_TYP_DESC`, the name of park and housing `PARKS_NM, HADEVELOPT`, the x and y coordinate for NY State Plane `X_COORD_CD, Y_COORD_CD`, and the latitutde and longitude `Latitude, Longitude`.

2.2 Data quality

First, we analyze the quality of data by each column using PySpark. Each PySpark job produces a key-value pair of [value of the cell, base data type (INT, TEXT, etc.), semi type (Location, time, etc.), quality label (Null, Valid, Invalid)], and for time-related features three extra value: [year, month, day] or [hour, minute ,second] for future modeling purpose. [Table 1] lists counts of nulls and invalid entries for each column.

Table 1: The number of nulls and invalid entries

Column Names	Number of Nulls	Number of Invalid Entries
CMPLNT_NUM	0	0
CMPLNT_FR_DT	655	8015
CMPLNT_FR_TM	48	903
CMPLNT_TO_DT	1391478	8015
CMPLNT_TO_TM	1387785	903
RPT_DT	0	0
KY_CD	0	0
OFNS_DESC	18840	0
PD_CD	4574	0
PD_DESC	4674	0
CRM_ATPT_CPTD_CD	7	0
LAW_CAT_CD	0	0
JURIS_DESC	0	0
BORO_NM	463	0
ADDR_PCT_CD	390	0
LOC_OF_OCCUR_DESC	1127341	0
PREM_TYP_DESC	33279	0
PARKS_NM	5093632	0
HADEVELOPT	4848026	0
X_COORD_CD	188146	0
Y_COORD_CD	188146	0
Latitude	188146	0
Longitude	188146	0

2.3 Detailed quality issues and cleaning

a. For date-related columns, we fill all null values with NULL. Since the dataset concludes data from 2006 to 2015 in its description, so we count all cells with dates ranged from 2006 to 2015 as valid. There is one cell with year value 1015, and we think this may be a typo, so we change it to 2015.

b. For time-related columns, there are 903 cells with hour value 24, which is invalid in 24-hour time frame. Commonly hour 24 refers to 00, so we change the cells with 24 hour value to 00.

c. i. For the column KY_CD, OFNS_DESC, the following pairs could be merged together because their descriptions are the same even though their codes are different. In addition, some codes have blank descriptions which might due to data entry error, so it is reasonable to merge them by same code number.

```
(('351', ''), 10)
(('351', 'CRIMINAL MISCHIEF & RELATED OF'), 433358)
(('121', 'CRIMINAL MISCHIEF & RELATED OF'), 72416)

(('116', ''), 2)
(('116', 'SEX CRIMES'), 10853)
(('233', 'SEX CRIMES'), 44117)
(('233', ''), 13)

(('232', ''), 1)
(('232', 'POSSESSION OF STOLEN PROPERTY'), 20376)
(('111', 'POSSESSION OF STOLEN PROPERTY'), 9112)
(('111', ''), 1)

(('234', 'PROSTITUTION & RELATED OFFENSES'), 111)
(('115', 'PROSTITUTION & RELATED OFFENSES'), 122)
(('356', 'PROSTITUTION & RELATED OFFENSES'), 609)
(('356', ''), 9)

(('232', ''), 1)
(('232', 'POSSESSION OF STOLEN PROPERTY'), 20376)
(('111', 'POSSESSION OF STOLEN PROPERTY'), 9112)
```

```
(('111', ''), 1)

(('234', 'PROSTITUTION & RELATED OFFENSES'), 111)

(('115', 'PROSTITUTION & RELATED OFFENSES'), 122)
(('356', 'PROSTITUTION & RELATED OFFENSES'), 609)
(('356', ''), 9)

(('676', 'NEW YORK CITY HEALTH CODE'), 31)
(('366', 'NEW YORK CITY HEALTH CODE'), 44)
(('366', ''), 1)

(('678', ''), 223)
(('678', 'MISCELLANEOUS PENAL LAW'), 8253)
(('126', 'MISCELLANEOUS PENAL LAW'), 110468)

(('122', 'GAMBLING'), 122)
(('350', 'GAMBLING'), 1901)

(('118', 'DANGEROUS WEAPONS'), 50563)
(('236', 'DANGEROUS WEAPONS'), 73672)

(('235', ''), 11)
(('235', 'DANGEROUS DRUGS'), 285790)
(('117', 'DANGEROUS DRUGS'), 62679)

(('118', 'DANGEROUS WEAPONS'), 50563)
(('236', 'DANGEROUS WEAPONS'), 73672)
(('236', ''), 8)
```

ii. For code 124, `KIDNAPPING, KIDNAPPING & RELATED OFFENSES, KIDNAPPING AND RELATED OFFENSES` looks like the same entry and are ready to be merged.

```
(('124', ''), 1)
(('124', 'KIDNAPPING'), 2)
(('124', 'KIDNAPPING & RELATED OFFENSES'), 2300)
```

```
(('124', 'KIDNAPPING AND RELATED OFFENSES'), 2)
```

iii. For code `685, 675, 365, ADMINISTRATIVE CODE, ADMINISTRATIVE CODES` might refer to the same entry and are ready to be merged.

```
(('675', ''), 4)
(('675', 'ADMINISTRATIVE CODE'), 1446)
(('365', 'ADMINISTRATIVE CODE'), 9937)
(('365', ''), 2)
(('685', 'ADMINISTRATIVE CODES'), 18)
```

iv. We merge `(('120', 'ENDAN WELFARE INCOMP'), 22)` and `(('345', 'ENDAN WELFARE INCOMP'), 122)` since they have the same descriptions.

v. The description of `(('120', 'CHILD ABANDONMENT/NON SUPPORT'), 367).` is not consistent with the more common descriptions associate with code `120: ENDAN WELFARE INCOMP`. This line may be wrongly recorded, so we keep this one separate and not merge it.

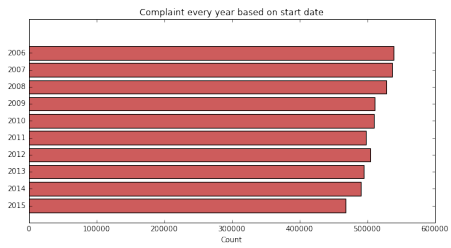
2.4 Data summary

We find following interesting distributions of several features, and these provide us a big picture of the crime data.

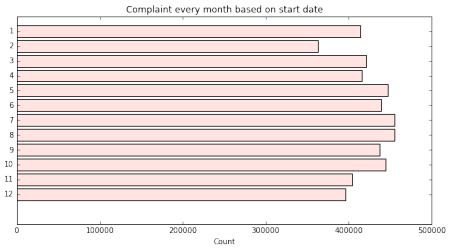
a. **Time** [Figure 1.]
We plot distribution by the exact year, month and hour of occurrence of crime. For the valid data, the number of crimes happened in NYC decreases every year; the crime rate is relatively higher during summer time, and it seems that crime is more likely to happen from afternoon till evening.

b. **Crime type** [Figure 2.]
We plot the top 20 most frequent crime type for last 10 years in NYC. It seems that assault, harassment and drugs are the most common crimes happened in town.

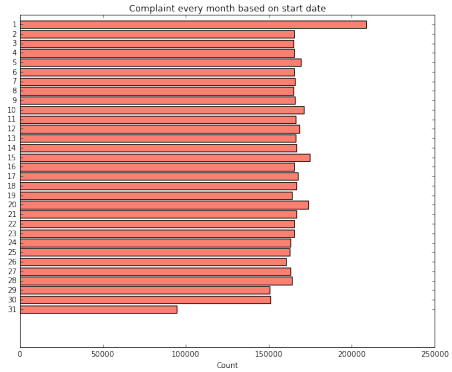
c. **Location** [Figure 3.]
Crimes happened in Manhattan and Brooklyn are slightly more than other other borough. More crimes happen in street or apartment, and we filter out the most commonly happened [Latitude, Longitude] for future analysis.



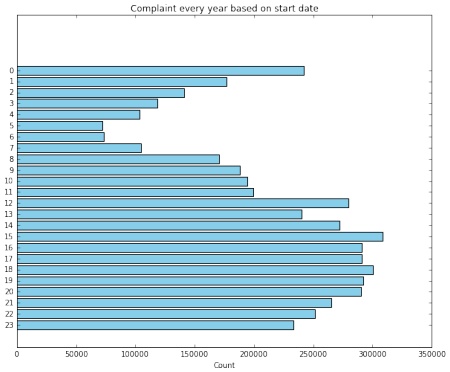
(a) By year from 2006 to 2015



(b) By month



(c) By day



(d) By hour

Figure 1: Illustration of distribution of exact date-time of occurrence of crime

3 PART II: DATA EXPLORATION

3.1 Experimental techniques and methods

Original NYPD Crime dataset: Of all features we find the following interesting and worth being examined: `[year, month, day, crime_code, crime_type, borough]`.

Additional external dataset: In addition to the NYPD crime dataset we used in Part I, we also scrap weather dataset and social behavior datasets including poverty rate (2005-2014), graduation rate (2006-2015), unemployment rate (2003-2015), drinking age(youth) (2005-2013, every two years), number of heavy drinkers seeking treatment each year (2005-2015) from various online sources. The original weather dataset contains features including temperature, visibility, wind speed and precipitation from 2006 to 2015. The poverty rate dataset contains data for each of five boroughs from year 2005 to 2014, but the poverty rates in Staten Island in 2006 and 2008 are missing.

In part I, we use mapper method to create key-value pairs for each column. In this part, we use reducer method in PySpark to count 5 dataframe we think relevant for future analysis

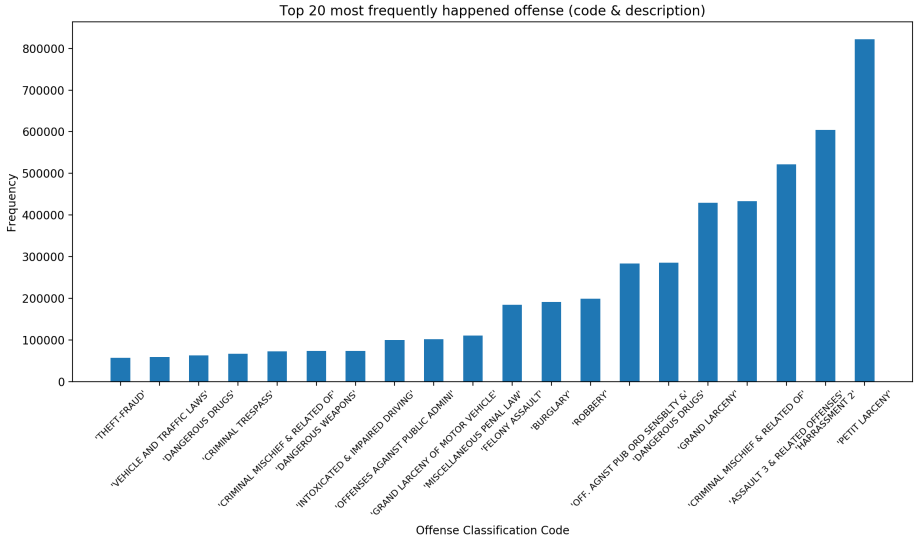
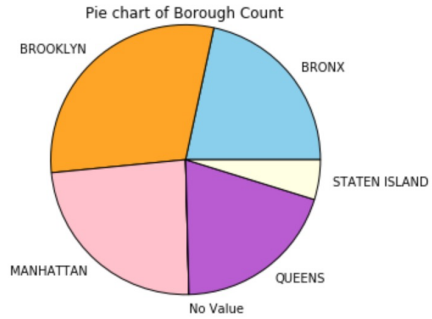


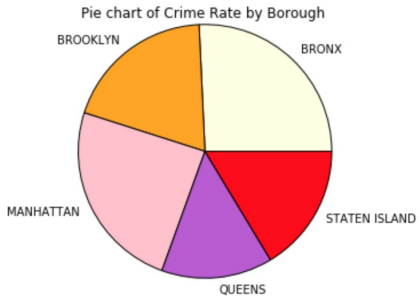
Figure 2: Top 20 most frequently happened offense

Table 2: PySpark output files format

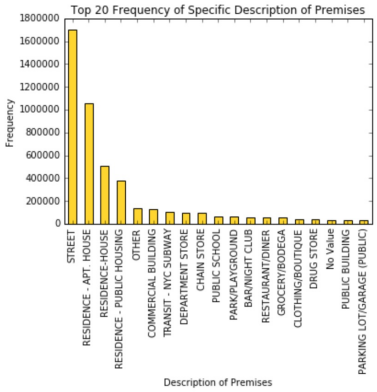
Value Pairs	**Count**
((year, borough), count)	50
((year, month, borough), count)	600
((year, monyh, crime_type, borough), count)	42000
((year, latitude, longitude), count)	1353924
((year, month, precinct), count)	9240



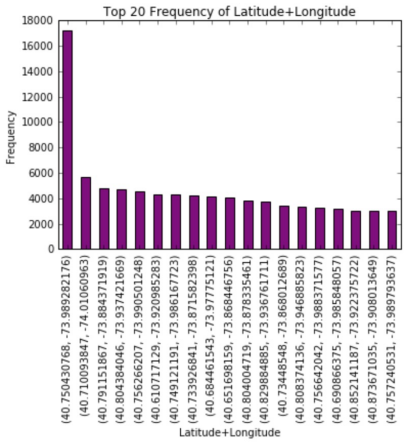
(a) By borough, count



(b) By borough, crime rate



(c) By location



(d) By exact coordinates

Figure 3: Illustration of distribution of location

[Table 2.]. We also run time series analysis on the crime count happened in the 5 borough of NY to give an overview of seasonal trend.

3.2 Hypothesis I: Poverty rate is positively correlated with crime count

The poverty rate dataset contains data for each of five boroughs from year 2005 to 2014, but the poverty rates in Staten Island in 2006 and 2008 are missing. We examine the time series results in total and by different borough: Our hypothesis is proven in Brooklyn, Manhattan, Bronx and Queens with a positive correlation, but for Staten Island the correlation is not significant, and our guess is that the data size is relatively small to make a strong statement. Alternatively, if we look at the poverty rate and crime count each year cross-borough, it also matches our hypothesis that the higher the poverty rate in that borough, the higher the crime count.

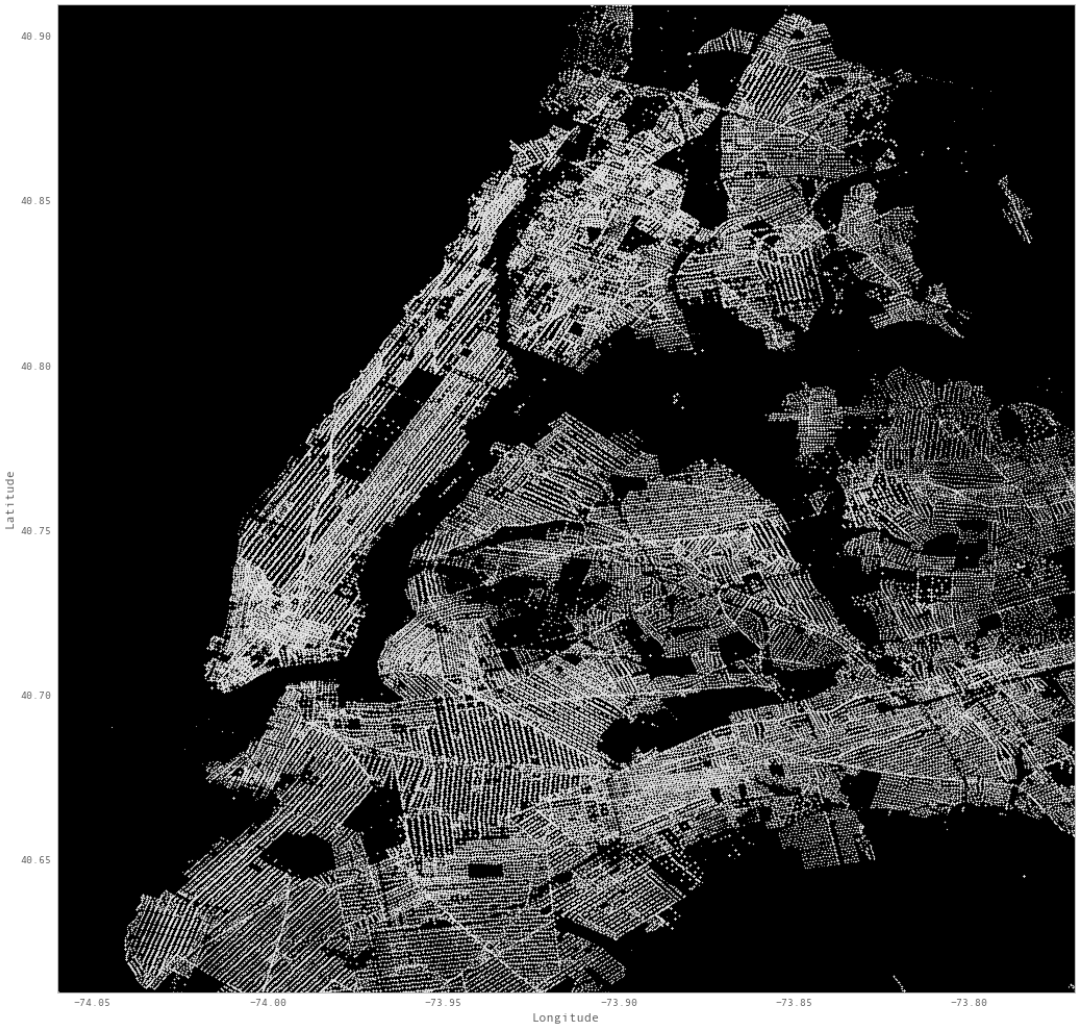
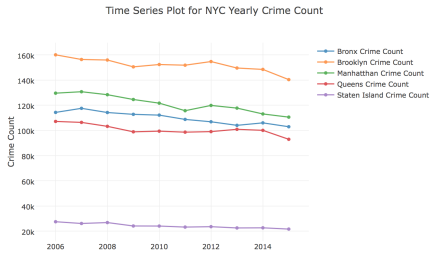
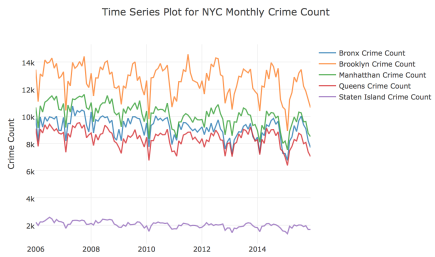


Figure 4: A map of nyc crime by latitude and longitude

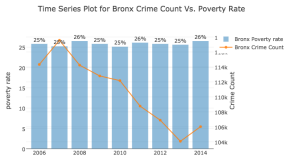


(a) Time series, by year

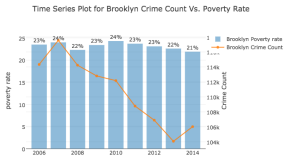


(b) Time series, by count

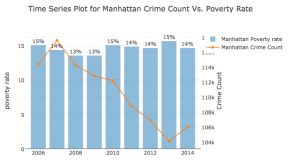
Figure 5: An overview of the trend of NY crime



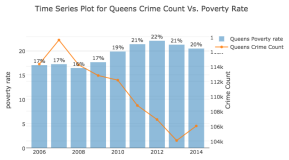
(a) Bronx



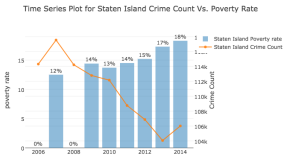
(b) Brooklyn



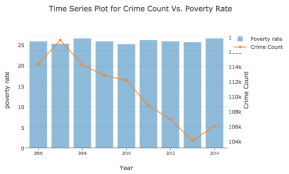
(c) Manhattan



(d) Queens



(e) Staten Island



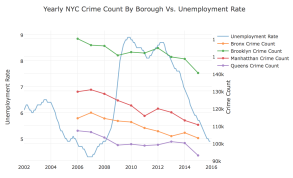
(f) Total

Figure 6: Time Series analysis for crime count v.s. poverty rate

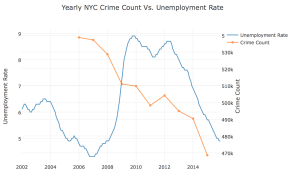
3.3 Hypothesis 2: Unemployment rate has a leading effect on crime count

We first examine the relationship between the total crime count and the unemployment rate. It has a clear positive leading effect on the crime count: the decreasing unemployment rate in 2002-2006 lead to the decreasing trend of crime count in 2006-2010; due to the financial crisis in 2008, the unemployment rate significantly increases and the crime count rises correspondingly in around 2010. This hypothesis can also be further emphasized by the time series analysis by borough, where the crime rate in Manhattan tremendously increases from 2010 to 2011, since financial crisis may have most obvious negative influence on Wall Street. The unemployment rate gradually decreases since 2012, and the crime rate consequently drops roughly from 2013, and we expect it to continuously drop in the future, given the

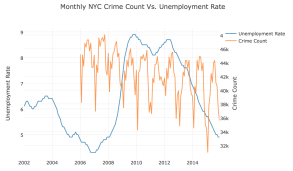
current stable economic situation.



(a) yearly count, by borough



(b) yearly count, total

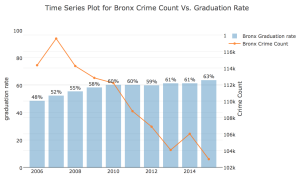


(c) monthly count, total

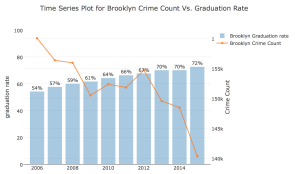
Figure 7: Time Series analysis for crime count v.s. unemployment rate

3.4 Hypothesis 3: Emphasis on education will decrease the crime rate

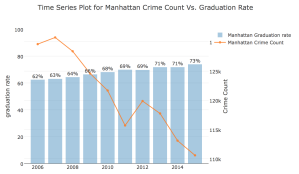
The time series and bubble chart state our hypothesis directly, that the higher the graduation rate, the lower the crime rate. It makes sense since eduation together with parenting, neighborhood, etc. does play an important role on crime, especially youth crime, and we explore some similar features in other hypothesis.



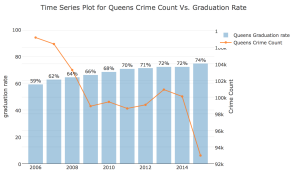
(a) Bronx



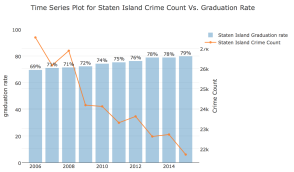
(b) Brooklyn



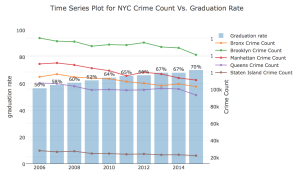
(c) Manhattan



(d) Queens



(e) Staten Island



(f) Total

Figure 8: Time Series analysis for crime count v.s. graduation rate

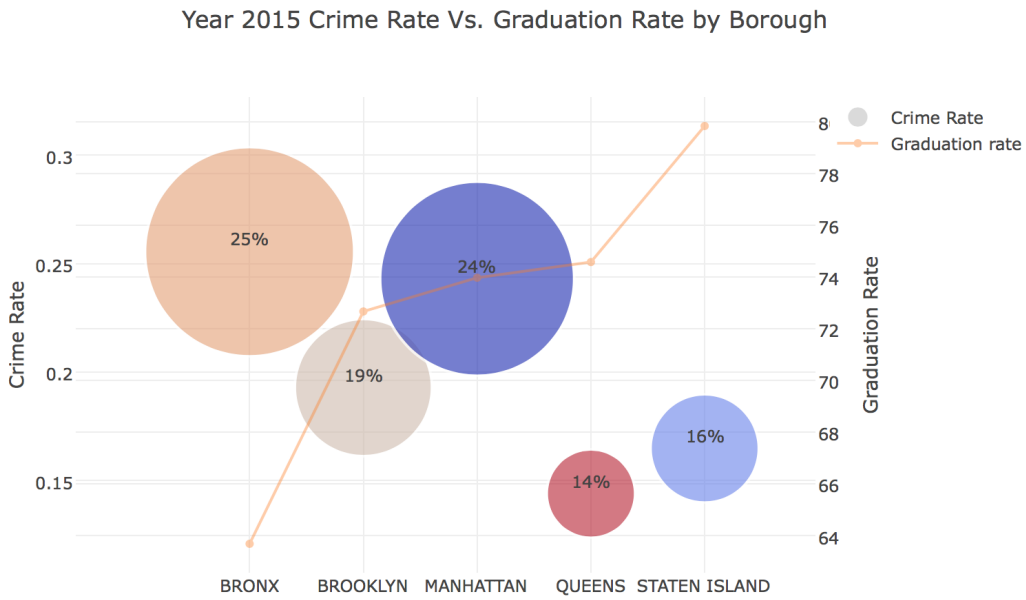


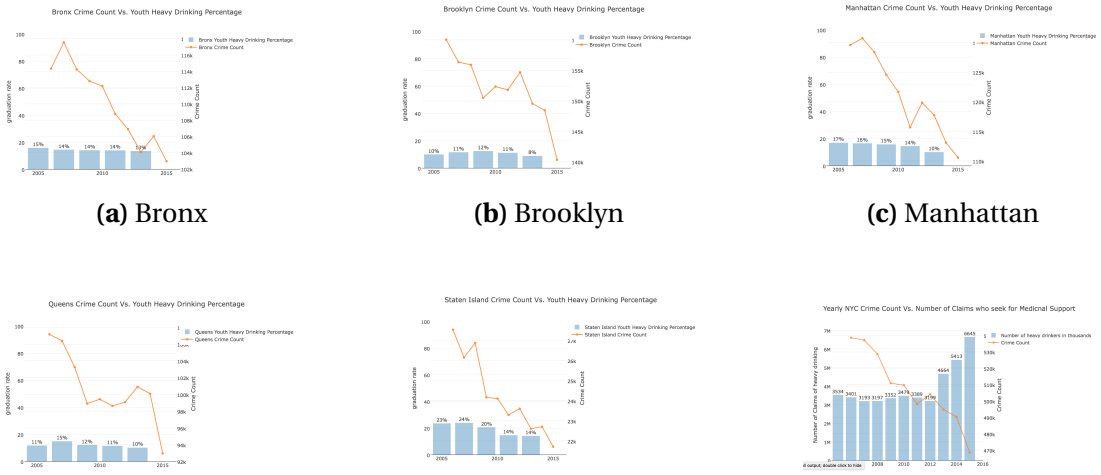
Figure 9: The graduation rate negatively correlates to crime counts

3.5 Hypothesis 4: High youth heavy drinking rate is a strong indicator for high crime rate

We analyze the effect of heavy drinking rate of both youth and adults, and find the youth heavy drinking rate has a significantly higher influence on the crime rate. We run the regression model and produces a very high correlation score (for example, corr=0.78 for Brooklyn). This feature may have a joint effect with the graduation rate and poverty rate that we talked before, as growing up in a bad environment, both in terms of family and community, can have a negative effect on a person's characteristic and influence his/her behavior.
Surprisingly, the heavy drinking rate of adult has a less significant impact on crime rate, and our guess is that adults are more responsible for his/her behavior under effect of drinking.

3.6 Hypothesis 5: The number of heavy drinkers seeking for medical support is negatively correlated to crime count

We run this analysis as a further examination of influence of heavy drinking on crime count and rate. The correlation is even more clear in this case: when there are more heavy drinkers seeking for medical support, the crime count is low. High number of claims may be an indicator that government is taking actions on heavy-drinking related problems, and therefore the crime rate decreases.



(a) Bronx **(b)** Brooklyn **(c)** Manhattan

(d) Queens **(e)** Staten Island **(f)** Total

Figure 10: Time Series analysis for crime count v.s. youth heavy drinking rate

3.7 Hypothesis 6: Temperature can influence the number of happening of some specific crime types

Data Preparation

From part I, we find that there are several crime type codes and its corresponding descriptions can be merged together, so we first use python to convert descriptions, which indicate the same crime type code, into one form. Then we append the average temperature for every month to the previous table and group the table by key using MySQL (crime type, temperature). The correlation between temperature and the number of occurrence of each crime type is 0.023888668794923695, which is not significant. Considering the temperature in each season does not change much, we decide to divide temperature into four bins, which indicate four seasons. (1: 1 3, 2: 4:6, 3: 7 9, 4:10 12). After recalculating the crime counts for each crime type in each season, the correlation between season and crime counts we get is 0.007164988065037319.

Detailed Analysis

From [Figure 12.], we can see that the number of occurrence of each crime type in season seems consistent every year, so we can use one-year data to further analyze on the correlation between temperature and the probability of occurrence of certain crime types. Top four most frequently happened crime types are petit larceny, harrassment, assault 3 & related offenses, and grand larceny [Figure 13.]. The correlation between temperature and each crime type is shown below [Table 3.], we can see that the correlation is very high and the p-value is very

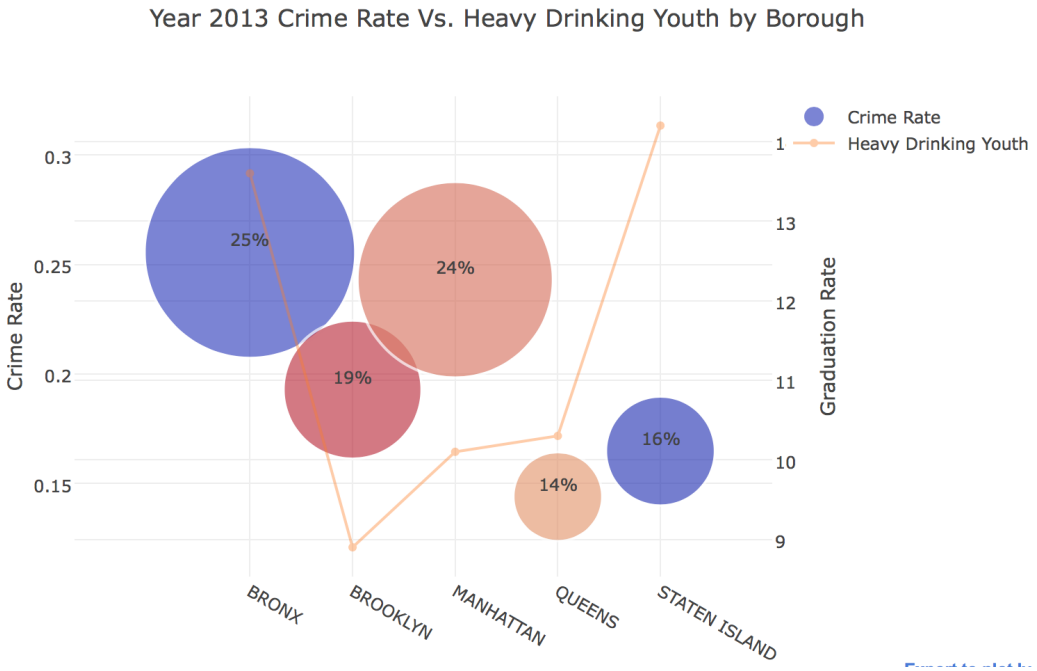


Figure 11: The heavy drinking rate v.s. crime rates, bubble charts

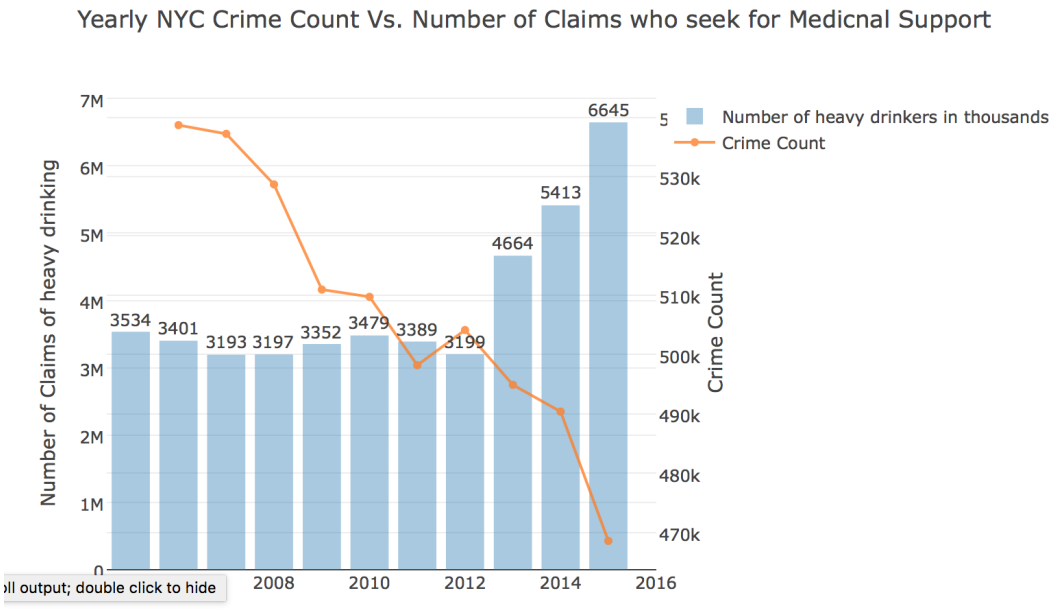


Figure 12: The number of heavy drinkers seeking for medical support is negatively correlated to crime count

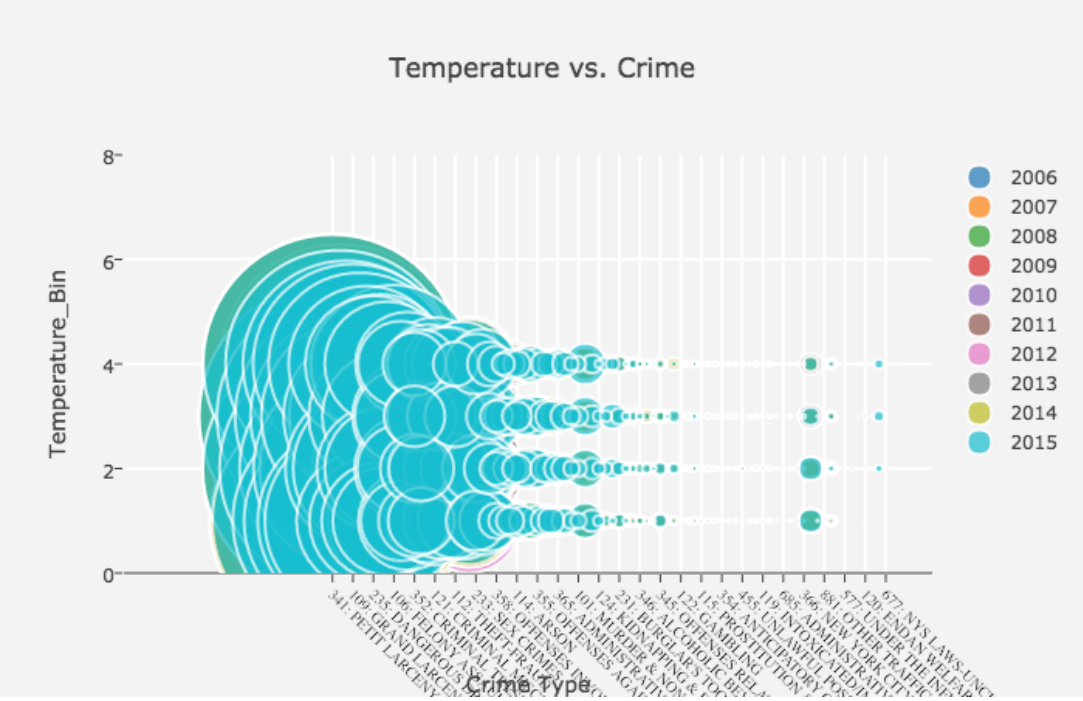


Figure 13: Time series bubble plot illustrating the season and crime type

Table 3: Correlation between crime type and temperature

Crime Type	Correlation	p-value
341: Petit larceny	0.95137	1.97346e-06
578: Harrassment 2	0.97451	8.11226e-08
344: Assault 3 & related offenses	0.92127	2.08529e-05
109: Grand larceny	0.89976	6.72244e-05

low. This result can verify our hypothesis, which is temperature has effect on crime type.

	crime	Temperature	Count
370	341: PETIT LARCENY	79.09677419	7660
371	341: PETIT LARCENY	79.22580645	7551
367	341: PETIT LARCENY	68.77419355	7432
368	341: PETIT LARCENY	71.5	7322
369	341: PETIT LARCENY	74.73333333	6982
366	341: PETIT LARCENY	58.29032258	6899
365	341: PETIT LARCENY	54.53333333	6709
364	341: PETIT LARCENY	52.93333333	6421
363	341: PETIT LARCENY	51.09677419	6359
362	341: PETIT LARCENY	38.35483871	6215
681	578: HARRASSMENT 2	79.09677419	5809
361	341: PETIT LARCENY	30.16129032	5806
682	578: HARRASSMENT 2	79.22580645	5782
678	578: HARRASSMENT 2	68.77419355	5659
680	578: HARRASSMENT 2	74.73333333	5576
679	578: HARRASSMENT 2	71.5	5503
677	578: HARRASSMENT 2	58.29032258	5363
415	344: ASSAULT 3 & RELATED OFFENSES	68.77419355	4967
419	344: ASSAULT 3 & RELATED OFFENSES	79.22580645	4962
675	578: HARRASSMENT 2	52.93333333	4954
676	578: HARRASSMENT 2	54.53333333	4913
360	341: PETIT LARCENY	24.17857143	4849
418	344: ASSAULT 3 & RELATED OFFENSES	79.09677419	4849
674	578: HARRASSMENT 2	51.09677419	4796
673	578: HARRASSMENT 2	38.35483871	4761
416	344: ASSAULT 3 & RELATED OFFENSES	71.5	4587
417	344: ASSAULT 3 & RELATED OFFENSES	74.73333333	4495
672	578: HARRASSMENT 2	30.16129032	4299
414	344: ASSAULT 3 & RELATED OFFENSES	58.29032258	4274
410	344: ASSAULT 3 & RELATED OFFENSES	38.35483871	4182
413	344: ASSAULT 3 & RELATED OFFENSES	54.53333333	4163
411	344: ASSAULT 3 & RELATED OFFENSES	51.09677419	4119
412	344: ASSAULT 3 & RELATED OFFENSES	52.93333333	4058
74	109: GRAND LARCENY	79.22580645	3967
73	109: GRAND LARCENY	79.09677419	3951
69	109: GRAND LARCENY	58.29032258	3925

Figure 14: Top four most frequently happened crime type and corresponding temperature by count

Table 4: Contribution

Group Member	Contribution
Yichao Chen	wrote scripts summarizing data in columns 1, 7-13 and plot distribution plots for each column. Run analysis on weather data.
Zoe Ma	wrote scripts summarizing data in columns 2-6 and plot distribution of related data. Deal with all date-time related data and run regression.
Yanjia Zhang	wrote scripts summarizing data in columns 14-24 and drawing plots to visualize each column. Produce time series results.

Table 5: Correlation table between external features and crime

External Dataset v.s. Crime Count	Weather	Poverty Rate	Unemployment Rate	Graduation Rate	Drinking Age (Youth)	Health Charity for Alcoholic
Correlation	0.02388	0.56006	0.17671	-0.84481	0.78612	-0.80323

4 CONTRIBUTION

[Table 4.]

5 CONCLUSION

Based on the correlation results on the six external features, we can see a strong correlation between crime rate and graduation rate, youth heavy drinking, and heavy drinkers seeking help. The correlation value is relatively small between poverty rate and crime, but the p-value=6.32175e-05 is very small, which indicates a significant correlation as well. The unemployment rate has a leading effect on crime, and from [Table 3.] for top recorded crime types, the temperature also has a strong causal effect.

Further on our analysis results, we think the poverty rate, graduation rate, and youth heavy drinking have a joint effect on crime, as it indicates the influence of community and family environment on criminals. Broken families, family criminal background, low school adjustment can all be factors leading to juvenile crime[8], which is captured by our analysis on youth heavy drinking and graduation rate. The effect of unemployment rate is lagged a bit to show on crime. For the weather, In the future, we can stratify crime types according to the degree of the violence, and then we can test on correlation between the degree of violence and temperature.

REFERENCES

Dataset:
1. NYC Population Data By Borough
https://www.citypopulation.de/php/usa-newyorkcity.php
2. NYC Unemployment Statistics
https://www.bls.gov/lau/data.htm
3. NYC Youth Heavy Drinking Data
http://a816-dohbesp.nyc.gov/IndicatorPublic/VisualizationData.aspx?id=2050,4466a0,97,Summarize
4. NYC Graduation Rate
http://schools.nyc.gov/Accountability/data/GraduationDropoutReports/default.htm
5. Nyc Weather Data https://www.wunderground.com/history/
6. Nyc Poverty Rate By Borough
http://www.nyc.gov/html/ceo/html/poverty/lookup.shtml
7. NYC Medical Support Summary
https://data.ny.gov/Human-Services/OASAS-Medicaid-Trend-Recipient-Summary-Profile-Beg/g4vm-hyyi/data
Paper:
8. Chris Knoester and Dana L. Haynie, ?Community Context, Social Integration Into Family, and Youth Violence,? Journal of Marriage and Family 67, no. 3 (2005): 767-780.